FRONTIER DEVELOPMENT LAB 2018: ASTROBIOLOGY TEAM 2

From Biohints to Confirmed Evidence of Life: Possible Metabolisms Within Extraterrestrial Environmental Substrates

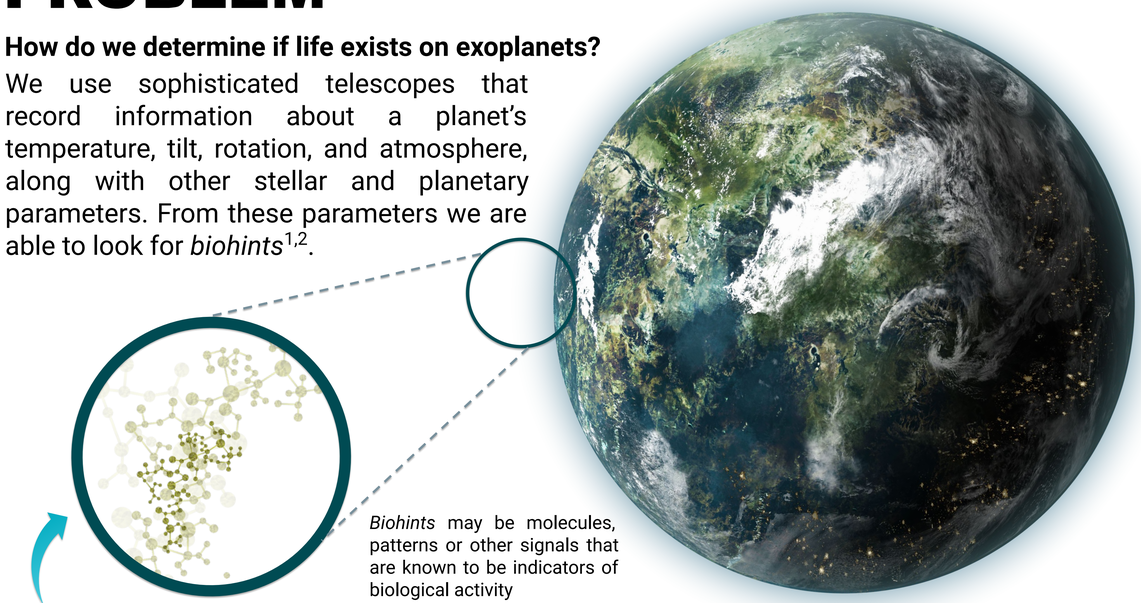
Frank Soboczenski[3], Michael D. Himes[1], Molly D. O'Beirne[2], Simòné Zorzan[4], Atılım Güneş Baydin[5], Adam Cobb[5], Yarin Gal, Massimo Mascaro, Daniel Angerhausen[7], Geronimo Villanueva, Shawn D. Domagal-Goldman[6] and Giada N. Arney[6]

Google Cloud

PROBLEM

How do we determine if life exists on exoplanets?

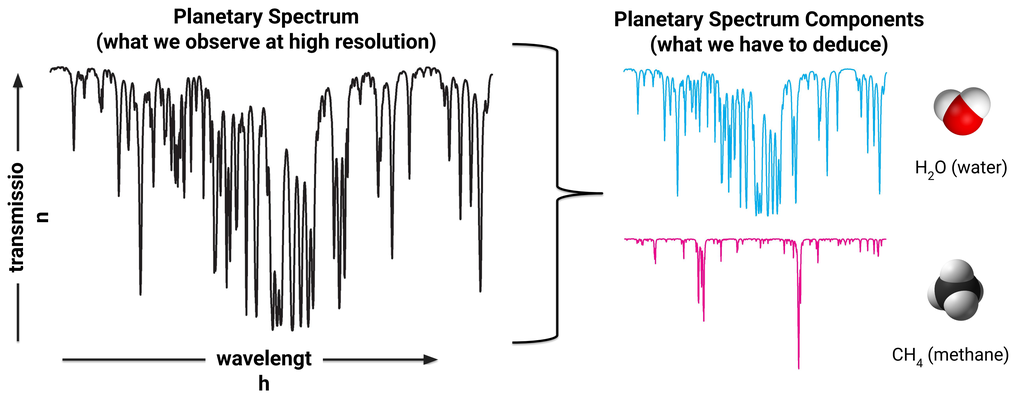
We use sophisticated telescopes that record information about a planet's temperature, tilt, rotation, and atmosphere, along with other stellar and planetary parameters. From these parameters we are able to look for *biohints*[1,2].



Biohints may be molecules, patterns or other signals that are known to be indicators of biological activity

We want to know what molecules are in the atmosphere of an exoplanet.

Knowing this can help us determine whether or not life may exist on an exoplanet. This is because certain combinations of molecules are indicative of life[1,2].



What we are able to observe is <u>complicated.</u>

Telescopes record emissions from molecules in a planet's atmosphere at different wavelengths. This results in a complicated planetary spectrum, which we then have to deconvolve into potential atmospheric molecular components. This process (called an atmospheric retrieval) is very time-consuming and computationally expensive!

Can we use machine learning to expedite the speed and accuracy of determining the composition of exoplanetary atmospheres?

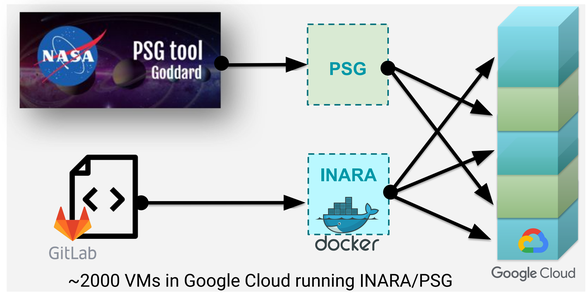
SOLUTION

INARA: Intelligent exoplaNet Atmospheric RetrievAl

Uh, Houston, we need data!

3 million synthetic planetary spectra were generated using PSG (Planetary Spectrum Generator[3], courtesy of Geronimo Villanueva at NASA Goddard) and compute resources supplied by Google Cloud.

High resolution spectra were generated over a range of stellar and planetary parameters (28 total) to maximize the diversity of the produced dataset for machine learning and release to the scientific community.



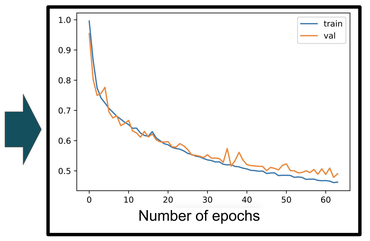
~2000 VMs in Google Cloud running INARA/PSG

Proof of Concept: Synthetic Spectra Input

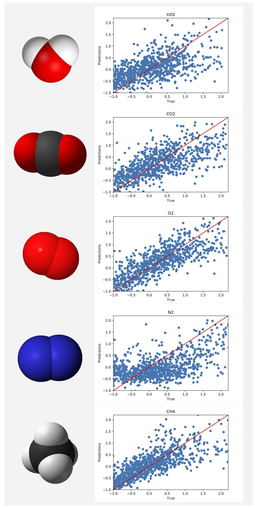
Set	Current	Future
Training	100,000	2.5 million
Validation	10,000	400,000
Test	7,710	200,000

Machine Learning Models

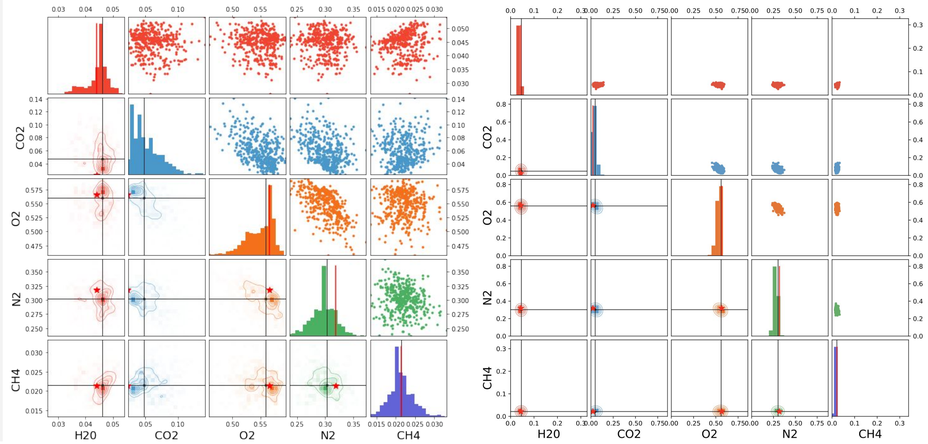
We explored many model architectures ranging in complexity from linear regression and feed-forward neural networks to convolutional neural networks (CNNs). We present results from the best performing model, a 1D CNN with the following configuration: Conv1d(64) - tanh - MaxPool - Conv1d(64) - relu - MaxPool - Conv1d(128) - relu - MaxPool - Conv1d(256) - relu - FC(256) - relu - FC(12) - T. Loss(0.42) - V. Loss (0.49) - 64 epochs.



Loss plot for the chosen CNN



True vs CNN predicted values



*Posterior distributions of the relative molecular abundances for one planet (600 predictions for each molecule). Within each scatter plot, each plot is a single regression in the CNN. The straight lines indicate the median values and the star indicates the true value (**right:** high level overview - **left:** zoomed in scale)*

Comparison

Method	Time	Molecules retrieved	Error	H_2O	CO_2	O_2	N_2	CH_4
Traditional	Hours to days	User-specified						
ExoGAN[4]	Minutes	H_2O, CO, CO_2, CH_4	MSE	3.43e-4	1.02e-2	7.00e-3	2.05e-2	1.93e-4
HELA[5]	Seconds	H_2O, HCN, C_2H_2						
INARA	Seconds	H_2O, CO, CO_2, CH_4, C_2H_6, O_2, O_3, N_2, N_2O, NO_2, NH_3, SO_2	± 2σ	2.28e-3	3.53e-2	2.59e-2	5.21e-2	1.07e-3